DOCUMENT RESUME

ED 150 331 CE 014 520

AUTHOR West, Leonard J.
TITLE Design and Conduct of Educational Surveys and Experiments. Delta Pi Epsilon Service Bulletin No. 2.
INSTITUTION Delta Pi Epsilon, St. Peter, Minn.
PUB DATE Jan 77
NOTE 41p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Business Education; *Experiments; Research; *Research Criteria; *Research Design; *Research Methodology; Research Skills; Sampling; Statistical Analysis; *Surveys; Validity

ABSTRACT

This bulletin for the researcher, the doctoral or master's candidate, or teacher interested in and/or conducting research is intended to (1) alert novice investigators to the major requirements and pitfalls of survey and experimental research, (2) point out the weaknesses of past business education studies, and (3) provide consumers of research with some basis for evaluating research reports and claims. Four sections are included, with most of the content being presented in the first two sections. The first section on surveys covers questions amenable to survey methods, estimation of population parameters from sample statistics, accessibility, principal features of sampling, nonresponse, reliability of survey responses, comparing samples with population distributions, and phrasing of survey questions. It concludes with a summary of major survey requirements. The second section on experiments discusses the definition, purposes, and criteria of experimentation; four weaknesses of past business education experiments; major features of experimentation (sampling, wording of hypotheses, controlling for irrelevant sources of variation, valid and reliable criterion measures, and statistical requirements); and replication as a basis for external validity. The section concludes with a summary of major experimental requirements. The last two sections briefly discuss the requirements for research reporting and the impact of research on practice. (EM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED150331

DESIGN AND CONDUCT OF EDUCATIONAL SURVEYS AND EXPERIMENTS

Leonard J. West

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY:

ΔΠΕ

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U S DEPARTMENT OF HEALTH EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Delta Pi Epsilon
Service Bulletin No. 2

CE 014 520

Preface

Delta Pi Epsilon launched the new planned publications program during the 1976-77 biennium with two series, one targeted at the researcher and the other, the classroom teacher.

The *Service Bulletin* series is designed for the researcher, the doctoral or master's candidate, or teacher interested in and/or conducting research. The *Rapid Reader* series is geared for the teacher to read and immediately take ideas into the classroom for enlivening and enriching the teaching of business subjects. Each publication in both series is written by a recognized authority in the field. Several are Delta Pi Epsilon Research Award or Robert E. Slaughter Research Award winners.

Delta Pi Epsilon is proud to present these new publications for use in classes such as methods of teaching, principles of business education, and research classes as well as for use of teachers who want to remain abreast of the latest teaching/learning strategies and research techniques.

Publications released during the 1976-77 biennium include:

Service Bulletin No. 1—*Research: Process and Product* by Mildred Hillestad, The Ohio State University, Columbus, Ohio

Service Bulletin No. 2—*Design and Conduct of Educational Surveys and Experiments* by Leonard J. West, Baruch College, The City University of New York, New York, New York

Rapid Reader No. 1—*Teaching General Business* by Betty Jean Brown, University of Tennessee, Knoxville, Tennessee

Rapid Reader No. 2—*Teaching Consumer Education* by David K. Graf, Northern Illinois University, DeKalb, Illinois

Information concerning the above publications and others to be issued in each series may be obtained from the National Office of the Fraternity.

©Copyright, 1977 by Delta Pi Epsilon. Printed in the U.S. of America

---

DPE NATIONAL EXECUTIVE BOARD:
National President: Wanda Blockhus, School of Business, San Jose State University, San Jose, CA 95192.
National Vice-President: Robert A. Ristau, College of Business, Eastern Michigan University, Ypsilanti, MI 48197.
National Secretary: John S. Nigro, North Haven High School, North Haven, CT 06473.
National Treasurer: Jeanne L. Holley, Box 24, Conner Hall, University of Mississippi, University, MS 38677.
National Historian: Robbie Woodall, Nashville State Technical Institute, Nashville, TN 37209.
National Past President: Gordon F. Culver, 300 Teachers College, University of Nebraska-Lincoln, Lincoln, NE 68588.
DPE NATIONAL PUBLICATIONS COMMITTEE:
Chairman: Richard J. Dallas, Santa Rosa Junior College, 1501 Mendocino Avenue, Santa Rosa, CA 95401.
Member: Charles R. Hopkins, 254 Peik Hall, University of Minnesota, Minneapolis, MN 55455.
Member: Oleen M. Henson, 203 Ritter Addition, Temple University, Philadelphia, PA 19122.
Member: Carmela C. Kingston, Division of Business, Trenton State College, Trenton, NJ 08625.
DPE NATIONAL OFFICE:
National Executive Secretary: Ellis J. Jones, Gustavus Adolphus College, St. Peter, MN 56082.

Table of Contents

Page

Preface ..... Inside front cover
Table of contents i
Author's note ii

SURVEYS 1

Questions amenable to survey methods 2
- Nontriviality 2
- Generalizability 2
- Estimation of relationships 2

Estimation of population parameters from sample statistics 3
- Operational definitions 4

Accessibility 4

Principal features of sampling 5
- Representativeness 5
- Stratification 5
- Sample size 6

Nonresponse 7
- Follow-up and characterization of nonrespondents 8
- Determining reasons for nonresponse 8

Reliability of survey responses 9

Comparing sample with population distributions 9

Phrasing of survey questions 10
- Overstructuring 10

Summary of major survey requirements 11

EXPERIMENTS 12

Definition, purposes, and criteria 12
- Terminology of experimentation 13
- Other verbal considerations 14
- Internal and external validity as criteria of experimentation 14

Four prevailing weaknesses of business education experiments 16
- Excessive length 16
- Significant differences 16
- One-teacher experiments 17
- Prediction of outcomes in advance 18

Major features of experimentation 19
- Sampling and randomization 19

Wording of hypotheses 21
Controlling for irrelevant sources of variation 21
Randomization 21
Experimental control 22
Counterbalancing 22
Statistical control 23
Monitoring 24

Valid and reliable criterion measures 25

A miscellany of other statistical aspects 25
Statistical and practical significance 26
Correlation coefficients 26
Sample size 26
Statistically significant values 27
Certain experimental designs 27
Intact groups 28

Replication as a basis for external validity 28
Pitfalls in research 29

Summary of major experimental requirements 30

REQUIREMENTS FOR RESEARCH REPORTING 32

IMPACT OF RESEARCH ON PRACTICE 33

REFERENCES 34

Author's Note

This Bulletin treats the highlights of survey and experimental research and makes particular mention of leading misunderstandings and weaknesses in business education studies. As such, it is intended both to alert novice investigators to the major research requirements and pitfalls and to provide "consumers" of research with some basis for evaluating research reports and research claims. The small number of References is deliberately confined mostly to major compendia and summaries to permit the interested reader to locate much information in few sources.

Baruch College
City University of New York
January 1977

Leonard J. West

Design and Conduct of Educational Surveys and Experiments[1]

The design and conduct of research includes a host of features reminiscent of the verse:

> Great fleas have little fleas
> upon their backs to bite 'em,
> And little fleas have lesser fleas,
> and so *ad infinitum*.

In parallel fashion, research has big wheels and little wheels and wheels within wheels, so to speak. This Bulletin considers some of those wheels, particularly as applied to **surveys** (because they have been the most common type of research in business education) and to **experiments** (because only they can answer the most important educational questions). The major requirements for each of the two types of research are described and illustrated (largely in the context of questionnaire surveys and classroom experiments), giving particular attention to those features judged to be among the chief weaknesses in past and current business education research. Finally, research reporting and the impact of research on instructional practices are considered briefly.

Surveys

The issues treated here are: (1) Questions amenable to survey methods, (2) Estimation of population parameters from sample statistics, (3) Accessibility of the population to the investigator, (4) Principal features of sampling, (5) Nonresponse, (6) Reliability of survey responses, (7) Comparing sample with population distributions, and (8) Phrasing of survey questions. The writer's "Survey of the Activities of Entry-Level Bookkeepers in Relation to the High School Bookkeeping Curriculum" (West, 1973) happens to have invoked most of the issues just enumerated, and some of its features will be mentioned for illustrative purposes.

---

[1]This Bulletin incorporates a number of substantial revisions of my article in the Spring 1975 *NABTE Review*, which, in turn, incorporated my talks at the National Research Conference held in conjunction with the annual meetings of the National Business Education Association in Chicago on February 20, 1974 and February 26, 1975. For a technical review of this monograph I wish to express my appreciation to my colleague at Baruch College, Professor Patricia Kay. For editorial assistance, I am indebted to Ms. Alice Gallagher, Director of the Division of Business and Office Education of Harcourt Brace Jovanovich, and to the Publications Committee of Delta Pi Epsilon.

Questions Amenable to Survey Methods

Many business education surveys have solicited personal opinions on issues that are not or should not be matters of opinion at all. Prime examples are the many inquiries into teachers' instructional practices or preferences—the implicit assumption being that whatever the majority of teachers thinks best is best. Thus, there might be a preference for 2-minute tests over 5-minute tests at some stage of training in some business subject or for a focus on shorthand theory in early stenographic instruction. Proper test length, however, is determined by the reliabilities of the two tests, not by majority vote; and test reliability is a matter for measurement, not opinion. Whether or not to focus on shorthand theory is a matter settled by experimentation, not surveys. Indeed, virtually all questions concerning the relative merits of instructional methods and materials are experimental, not survey, questions. The only defensible purpose of such opinion surveys is to use their findings as a basis for re-educating teachers when prevailing practices are not in accord with reputable research evidence.

The first and foremost survey rule is: **Do not solicit opinions on issues that are not matters of opinion**. Furthermore, should one decide to determine what the prevailing practices are, it is scientifically and logically impermissible to infer merit from frequency, from popularity. Three additional major considerations also govern survey topics and purposes: nontriviality, generalizability, and estimation of relationships.

Nontriviality. Who cares (except, perhaps, equipment manufacturers) whether desk calculator X is used more frequently than desk calculator Y if both operate in substantially the same way? Some business education surveys have included just such trivia, violating the requirement that only information of practical consequence should be sought. If some outcome cannot make any difference in terms of practical action, exclude the question from the survey.

Generalizability. Narrow, parochial topics, such as "A Follow-Up Study of the Graduates of Blank High School," are also to be avoided; for it is hard to conceive of outcomes that would apply to and interest anyone outside that particular school and its community. Such inquiries might be the basis for a college term paper, but not for a thesis or dissertation or journal article. Research (whether survey or any other kind) should supply findings that are generalizable, that can properly be applied to an entire class of situations or events, rather than ones narrowly confined to the particular persons or events involved in some local situation.

A distinction should be made, however, with program- or curriculum-evaluation studies. In them, one necessarily studies the particular schools using the programs under evaluation. Credible studies of that kind, however, go far beyond mere collections of opinions as criteria of program merit.

Estimation of Relationships. Also unacceptable are routine fact-gathering studies that ask the question "How many?" but, beyond getting mere

nose counts, do nothing of consequence with the answer.[2] The fundamental requirement is to examine relationships; for example: between income and voting behavior, between educational background and occupational success, between curriculum content and life activities, and so on. Surveys are formally defined as studies that use samples to determine the incidence, distribution, and interrelationships of variables in the population.

Estimation of Population Parameters from Sample Statistics

Statisticians distinguish between census surveys and sample surveys. In the former, every member of the population is accounted for, as in the federal decennial census of the American population or in an inquiry addressed to the faculty sponsor of every Delta Pi Epsilon chapter. Whenever information is sought from every member of a population, we have a census. And a "population" does not necessarily refer to people, but instead to observations (often, of some feature of persons). The heights of persons, for example, constitute a population of heights; their typing speeds make up a population of typing speeds; their attitudes toward China, a population of attitudes toward China.

Census surveys, however, are vastly outnumbered by sample surveys, partly because a census of any scope can be formidably expensive, but mostly because it is rarely necessary to canvass an entire population. Suppose, for example, that one wished to estimate[3] the mean or average height of adult Americans, to estimate, that is, the "parameter for height" in the population of all adult Americans. Measurement of a sample of adult

---

[2]Complaints about educational surveys abound. Herriott's comment (1969, p. 1400) is: "The hurried preparation of a questionnaire to be administered to a convenient group of pupils or teachers in order to explore some problem of purely local interest...may represent 'a survey' to educators, but it is not 'survey research' as the term is known to social scientists." Trow's remark (1967, p. 319) is: "The bulk of survey research in education has been little more than a matter of asking some people some questions, and reporting the distribution of their responses." Advances in the methods of survey research are attributable to sociologists, not to educators; and Babbie's text (1973) is one good source of information on modern methods; in lesser detail and applied to educational and psychological inquiries, see Kerlinger (1973).

[3]An "estimate" in statistics is **not** a guess, but an approximation of an unknown and often unknowable "true" value. In contrast to **counting**, which results in exact values (e.g., there are just 28 students in some class, not 26 or 29½ or any other number), **measurement** results in estimates that are subject to "error" (departure from "true" values) from various sources. For example, a person found to weigh 128½ pounds on an ordinary physician's scale might be found to weigh 128 pounds and 6½ ounces on a more sensitive scale and 128.43272 pounds on some imagined scale that is still more sensitive and finely calibrated. Or one might round to 42 wpm a typing speed originally computed as 42.2 wpm—and even the correctness of that figure would depend, among several factors, on the precision of the watch used for the timing and on the typist's reaction time to start and stop signals. The foregoing examples illustrate just a few of the sources of "error" in all estimates. The immense power of statistics lies precisely in providing statable limits of error and statable probabilities that "true" values lie within the stated limits.

Americans would provide a statistic, and that statistic would be used as an estimate of the population parameter. Statistics are sample values, parameters are population values; and the former serve as estimates of the latter. Summarizing: The feature of interest in the population is called a **parameter**; the estimate of the population parameter provided by a sample is a **statistic** (Harris, 1969, p. 1309). Other illustrative examples are: identifying desirable instructional materials for typists, based partly on determining the incidence of typing from longhand copy among all employed typists—estimated from its incidence among a sample of such typists; or determining the advisability of including trial balance preparation in high school bookkeeping instruction, based partly on the incidence of that activity among all bookkeepers—estimated from its incidence among a sample of such bookkeepers.

The purpose of surveys—indeed of most educational research— is the estimation of population parameters from sample statistics.

Operational Definitions. Scientific discourse requires precise definitions of terms, definitions that permit the unambiguous assignment of an event as fitting or not fitting the definition. For example, the "operation" or act of reading the height of a column of mercury in a thermometer as above or not above 98.6° "defines" whether or not a person has a fever. As in the bookkeeping survey (West, 1973), such definitions can be crucial. In it, large numbers of persons employed to process or maintain financial records were found to be engaged in activities that had little direct correspondence with what is taught in high school bookkeeping. Although some observers maintained that such persons were clerks, not bookkeepers, the operational definition of a "bookkeeper" in that study had to be "a person employed to process or maintain financial records"—regardless of the correspondence, or lack of it, between job duties and high school curriculum content. Similarly, an "entry-level" position was defined as one available to persons without previous job experience in the field, regardless of school background. Distinguishing such persons from more experienced ones rested on examination of each person's job history and of the employer's hiring requirements. That information was solicited expressly to provide, among other things, unambiguous operational criteria for distinguishing entry-level from more experienced bookkeeping employees.

Accessibility

However terms may be defined operationally, the population from which the sample is to be drawn must be accessible to the investigator. To secure information from people, you must know which people make up your population and you must be able to find them. It is useless, for example, to attempt a survey of graduates a number of years after they have completed school if you cannot determine their current name [illegible] (e.g., of formerly single, but now married, women). The same requirement applies to surveys of records; you have to determine their existence and

availability before you start. For the purposes of the bookkeeping survey, for example, a complete and current population file of all 197,565 New York City private employers was available from the New York State Department of Commerce. Anyone who plunges into a survey without first verifying the existence and availability of *all* members of the pertinent population (so that a representative sample can be selected) is doomed to failure at the outset.

Principal Features of Sampling

Every respectable course or textbook on statistics and research deals with sampling, and there are overviews by specialists—for example, articles by Harris, Herriott, and by Tatsuoka in the *Encyclopedia of Educational Research* (1969). Some of the major concepts and procedures are considered here in relation to the fundamental requirement that the sample be "representative" of the population, that its characteristics match those of the population in all respects pertinent to the purposes of the research.

Representativeness. Representativeness is not a measurable entity but, rather, a consequence of the method of drawing the sample, whatever the size of the sample. The only representative samples are those drawn **at random**, a process that gives every member of the population an equal chance to turn up in the sample. Nonrandom samples are biased samples that preclude the drawing of inferences about populations. Thus, "volunteer" samples (ones readily at hand and willing to cooperate with the investigator) could well be (and often are) biased and unrepresentative because they are not selected at random from an entire population—a matter considered in more detail in the later discussion of experiments.

Stratification. Assume a survey of schools whose outcomes are expected to vary with school size (for example, one would expect more varied curricula in the larger schools). If so, one would sample from the population of schools in proportion to their sizes. For example, if it were known in advance that one fifth of all schools had 250-499 students, then one fifth of the schools in one's sample should be drawn from schools of that size. The population would be said to be "stratified" by size and a "proportional" sample drawn. In some other inquiry with other purposes one might stratify by IQ or sex or by whatever one or more characteristics are known from previous research to be related to one's purposes. In surveys, in contrast to experiments, the purpose of stratification, as Tatsuoka (p. 480 in the *Encyclopedia*) has pointed out, is to make each stratum (i.e., each subgroup) more homogeneous than the population at large, thereby reducing what is called the sampling variance and increasing the reliability of the estimate of the population parameter.

As in the school-size instance, the population distributions for the strata of interest can sometimes be determined in advance. Suppose, for another example, that in some occupational survey one wishes to stratify on the

basis of "type of employer." Population information of that kind is often available from one or another government agency or publication. A case in point is Silverthorn's vocabulary study (see West, 1963, for sampling details). Silverthorn recognized in advance that beyond the words needed for ordinary communication, each type of business would also have its own specialized vocabulary. Recourse to U.S. Department of Labor data revealed that 17.4 percent of the nation's secretaries, stenographers, and typists were in "finance, insurance, and real estate." Therefore 17.4 percent of all the words tallied by Silverthorn were from communications generated by that "type of employer." That procedure, applied to the other types of employers as well, provides assurance that the Silverthorn word list neither over- nor underrepresents the specialized vocabularies associated with the various types of employers. His word list is "representative" of "type of employer" by virtue of the proportional, stratified sampling employed.

At other times, although pertinent population distributions might be determinable in advance, either one cannot sample by stratum in advance or one wishes instead to survey an entire population of manageable size. What must be done in such instances to determine whether one has a representative sample is illustrated in the later section on "Comparing Sample with Population Distributions."

The bases for stratification vary with one's research purposes. In some inquiries, the pertinent population characteristic for sampling purposes might be sex or age or school background or income—to name just a few. The requirement, however, is that the basis for sampling have a known and nontrivial relationship to one's purposes. To invent an absurd instance of irrelevance, we do not sample by color of hair in instructional research because instructional outcomes are unaffected by color of hair. As will later be suggested, stratification by sex—however common in some educational research—is also often irrelevant and for that reason inappropriate. On the other hand, because in many areas of work men and women are not yet paid equally for equal work, job-income studies require stratification by sex. Similarly, studies of voting behavior necessarily stratify by socioeconomic status (among several factors or variables) because voting behavior is known to be influenced by such variables.

Concerning the drawing of stratified samples, then, one identifies the factors or variables that are known from prior evidence to be nontrivially related to outcomes and one stratifies on those bases. The result is statistics from sample surveys that provide more reliable estimates of population parameters *because* such sampling contributes to representativeness.

Sample Size. Sample surveys are intended to supply estimates of population parameters, and **the reliability of any estimate increases with sample size.** In sample surveys, then, the more the merrier. At the same time, it is wasteful to add to sample size when additions to the sample no longer significantly affect the findings. Silverthorn's vocabulary study is again illustrative. Blocks of 10,000 words were added cumulatively to the initial

40,000-word sample until the rank order for frequency of the 100 most common words was no longer changed. Because the rank order for 300,000 words did not differ from the one for 290,000 words, sampling stopped at 300,000 words examined. There are still other tactics for estimating optimum sample size, but they cannot be detailed here.

Representativeness has earlier been described as dependent on method of sampling and as requiring random sampling. Sample size is also a factor, however. When a sample is very small in relation to an extremely large and diverse population, it is nearly impossible to represent all the pertinent population characteristics in that sample. The result is potentially large error in the sample estimates of population values. A prominent example is the variety of outcomes of the several studies of the characteristics of business correspondence and papers. Tens of millions of business papers are spewed out *hourly* by the millions of employers across the nation; yet the several studies of business papers were based on samples numbering in the mere hundreds up to a few thousand business papers collected from a few to a few hundred employers. Small wonder, then, that the outcomes of such studies vary widely. The best that can be said is that the "truth" probably lies somewhere within the wide range of outcomes of the several studies.

Nonresponse

Nonresponse characterizes nearly all sample surveys, and the failure to deal with it properly is a flaw that is fatal; all findings become suspect because the representativeness of the sample is in question. Cooperation from all members of a sample of nontrivial size is unheard of. Easily the most compelling generalization that has arisen from the entire history of survey research in all fields is that nonrespondents tend to differ from respondents in ways material to the purposes of the survey. For example, persons who are ashamed of their low incomes (or who prefer not to make their high incomes known) tend not to respond to inquiries that include a question about income. Similarly, if your inquiry happens to include a question about age, you will not hear from persons who are reluctant to report their ages. In such instances, matters of interest that are related to income or age will be missing for persons at the extremes of the income or age distributions. Such samples are unrepresentative because they do not fully reflect the characteristics of a population that indeed includes persons at the extremes.

For reasons such as those illustrated, it is vital to determine whether nonrespondents differ significantly from respondents. Otherwise, the reliability of the findings is uncertain because the representativeness of the sample is uncertain. In fact, the frequency with which nonrespondents are found to differ from those who cooperate often justifies the assumption of low reliability in studies that ignore the problem of nonresponse. It is therefore sensible to assume that there will be nonrespondents and essential to plan in advance and in detail just what tactics will be employed to determine

whether significant discrepancies exist between respondents and nonrespondents. Very few business education (or other) surveys known to this writer have dealt with the problem of nonresponse; therefore, few of them permit high confidence in the reliability of their findings.

Follow-up and Characterization of Nonrespondents. A leading tactic that applies whenever responses have been received over a considerable time period (during which previous nonrespondents have been followed up a number of times) is to compare the early replies with the later ones. The tardy respondents who have had to be nagged several times before answering can be taken to typify those who never respond at all. If the responses to selected key questions from tardy respondents do not differ significantly from those of prompt respondents, it may be assumed that neither would the responses of those never heard from (see West, 1963, for a detailed illustration). It may then be inferred that, despite nonresponse, one's respondents are representative of the population.

It is important to recognize that the concern is with material, not trivial, reasons for nonresponse. Trivial reasons for initial nonresponse (on vacation, ill, busy, forgetful) suggest that it is wise to plan on several follow-ups, not merely one follow-up shortly after the initial contact. Classifying a tardy respondent as not unlike a nonrespondent is increasingly persuasive if that tardy respondent has had to be followed up a number of times.

Determining Reasons for Nonresponse. Try to determine the reasons for nonresponse in order to assess whether they are material to the purposes of the survey. In the bookkeeping survey (West, 1973), which was almost entirely confined to New York City firms, uncooperative employers and employees were telephoned. That telephoning had been planned in advance, in anticipation of an unpredictable proportion of nonresponse to the initial mail contact and expressly to determine the reasons for nonresponse. As it turned out, none of the reasons for nonresponse had anything to do with job duties—and job duties were the central focus and purpose of the survey. It was therefore reasonable to conclude that the job duties of nonrespondents, had they been determinable, would not have affected the findings based on those who did respond.

It must be emphasized, however, that these various tactics (comparison of prompt with tardy respondents, determining reasons for nonresponse) are not a fully adequate substitute for a very high rate of cooperation or response. Some authorities assert that anything less than an 80 or 90 percent response rate precludes definitive findings. Others are less concerned with rate of response and accept as sufficient evidence for representativeness nonsignificant differences between early and late respondents and nonmaterial reasons for nonresponse. On the latter score, the stated reasons may not be the real ones, especially when questions relate to a person's pride or self-image. Few will admit to engaging in forbidden activities or to not engaging in favored activities; similarly, people are reluctant to reveal limited success on such indices of success as income, school completion or, in the case of teachers, student learning. Thus, it is desirable to formulate

a few probing questions that test the reliability of reasons for nonresponse without tipping off or offending the prospective respondent. Formulating such questions is admittedly not easy.

Reliability of Survey Responses

As a check on the reliability of responses and the clarity of the questions, it is sometimes desirable to build redundancy into the survey instrument: by using more than one way to secure the information of exceptional interest. The bookkeeping survey, for example, inquired into the number of money columns in journals. The respondent who reported elsewhere in the questionnaire that he or she records incoming checks in a cash journal should also have been able to specify the number of money columns in that journal. When the latter information was not given, the respondent was contacted and asked to explain the discrepancy.[4]

Comparing Sample with Population Distributions

A survey sample may be drawn at random from within known-in-advance population distributions, or, sometimes, an entire population of manageable size might be surveyed. In either instance, if not all persons respond, it is necessary to determine whether respondent distributions match the known population distributions on the bases for stratification. Weeks, for example, surveyed all 463 "major market employers" in the Syracuse area (see West, 1963, for details), chiefly to determine their employment needs. Although responses were received from only 234 (50.5%) of those employers, Weeks was able to show (by statistical test) that there was no significant difference in the distributions for "type of firm" as between the 234 respondent firms and the population of all 463 firms. In the same fashion, if graduates of a collegiate school of business over a ten-year period are surveyed on matters that might be expected to be related to the length of their employment history since graduation, one would need to determine whether the distribution of graduation years among respondents differs significantly from the distribution for all graduates. For example, if registrar's or alumni office records show that 14 percent of all graduates during the period 1966-1975 were graduated in 1971, are very nearly 14 percent of the respondents 1971 graduates? If 64 percent of all graduates are males, do males make up approximately 64 percent of the respondents?

---

[4] Among 597 respondents there were several hundred such discrepancies distributed across six journals. All the discrepancies were cleared up by phone and turned out, for the most part, to reflect the great differences between the manual concepts and record forms of high school instruction and those of the computerized systems that prevail among the large employers who account for most of the employment in large cities.

Chi square is the usual, but not the only, statistical test for such comparisons, and it can be carried out in a few minutes on a desk calculator in rote fashion (as illustrated in West, 1963, for Weeks' data on type of firm). Whatever the bases for stratification and whatever the statistical test, it is mandatory for respondent and population distributions to be compared.[5] If statistically significant differences result, the representativeness of the respondent sample is in question. For that reason, so is the applicability or generalizability of the sample findings to the population.[6]

Phrasing of Survey Questions

Phrasing of survey questions to insure that they are free of ambiguity is not always easy. (In response to the test item "Columbus discovered America in _____," one fourth-grader wrote in the blank, "a boat.") Pretesting of data collection instruments is mandatory. Even with pretesting, some things can slip by. In the bookkeeping survey, for example, respondents were asked whether they had been promoted; but some took promotion to mean a raise in pay, whether or not accompanied by more responsible job duties. (The question "Have you been promoted since...?" should have been worded "Have you been promoted to more responsible job duties since...?") Age, as mentioned earlier, is a prominent example of sometimes vital information that some persons are reluctant to report. In the bookkeeping survey we asked for high school graduation date and, assuming graduation typically at age 18, were able to determine current age.

Overstructuring. Open-ended questions resist reliable interpretation of responses for coding and subsequent analysis. High structure (e.g., multiple choice, one- or two-word fill-ins) has the immense virtue of being self-coding or easily coded for analysis with near-perfect reliability. But for elite, sophisticated audiences, high structure—as Herriott has observed in his *Encyclopedia* article (1969, p. 1401)—can overabstract, oversimplify, oversummarize the issue under examination, thereby anatagonizing sophisticated respondents, giving them a low regard for the investigator and lead-

---

[5] See West (1963) for illustrations in business education settings of three sampling situations: (a) The distribution of the population characteristic is known, and one can guarantee in advance the drawing of a matching sample; (b) The population distribution is known in advance, but one cannot guarantee the drawing of a matching sample in advance; (c) The distribution of the characteristic in the population is not known in advance.

[6] Discrepancies affect generalizations across a population, not subgroups from that population. For example, if the proportion of small schools among respondents differs from the proportion of small schools in the population of all schools, one may not make generalizations about schools of all sizes, but only about schools of each particular size. For another example, if small-firm bookkeepers are disproportionately represented in a survey aimed at firms of all sizes, valid inferences can still be drawn about small-firm bookkeepers, but not about all bookkeepers in firms of all sizes (on matters other than size that are related to size).

ing to refusal to respond to the inquiry.[7] When information is sought from sophisticated audiences on debatable matters, it is desirable to conduct a pilot study using open-ended questions. Responses, after appropriate editing, can then be incorporated in structured form in the final questionnaire. Even so, it is wise to provide for an open-ended "Other (please describe)" response.

Summary of Major Survey Requirements[8]

Selection of a worthwhile (nontrivial, generalizable) topic—one that does not solicit opinions on issues that are not matters of opinion— and the use of findings to uncover relationships are evident requirements. There are no rule books for knowing what relationships to look for. That is an art requiring intelligence, experience, and close, sophisticated knowledge of the field under inquiry. Indeed, the relationships to be examined comprise the purposes of one's work and are specified in advance; one does not collect data at random in the hope that something of interest will be found therein.

Other major requirements are:

1. The target population has to be fully accessible.

2. If there are population characteristics known to be related to one's purposes, and if the population distributions on those characteristics are known in advance, sampling in proportion to those characteristics will increase the reliability of estimates of population parameters.

3. Whether across a population or from within strata, sampling must be at random.

4. Because reliability increases with sample size, the larger the sample the better—but short of increases in sample size that do not significantly affect the findings.

5. Data collection instruments should be pretested to insure their freedom from ambiguity. Highly structured questions facilitate high-reliability coding or interpretation of responses; but on opinion-dependent issues addressed to sophisticated audiences, overstructuring should be avoided. Also, as a check on the reliability of responses to the most crucial questions, redundancy can sometimes be built in by asking essentially the same question in a different way or by soliciting other cross-check information.

6. Response rates should be maximized by at least several follow-ups, if necessary.

---

[7]One instance is the soliciting of teachers' instructional preferences in a multiple-choice format. Unless a "None of the above" or "Other (please describe)" option is included, the respondent who does not favor any of the other options has no way to respond to the question.

[8]For the highlights and essentials of numerous types of research (survey, experimental, correlational, et al.), see Isaac and Michael's *Handbook in Research and Evaluation* (1971). Also see Kerlinger's article on "Research in Education" and Mouly's article on "Research Methods" (in the *Encyclopedia*, 1969).

7. Short of very high response rates, the problem of nonresponse must be dealt with, not ignored. On the crucial survey items, test for significant differences between the responses of very early and very tardy respondents and determine, if you can, the reasons for nonresponse.

8. If, despite follow-up, not all persons respond (assuming a stratified sample from a population whose distributions on the bases for stratification are known in advance), test for significant differences between the respondent and population distributions on each of the bases for stratification. Such tests determine whether respondents are representative of the population and, thereby, whether one's sample findings permit the drawing of generalizations about the population.

Experiments

Considered here, in turn, are: (1) Definition, purposes, and criteria of experimentation; (2) Four weaknesses of past business education experiments (excessive length, failure to find significant differences, one-teacher experiments, bases for hypotheses and prediction of outcomes); (3) Major features of experimentation (sampling, wording of hypotheses, controlling for irrelevant sources of variation, valid and reliable criterion measures, statistical requirements); and (4) Replication as a basis for external validity.

Definition, Purposes, and Criteria

While all experiments are research (properly accented on the second syllable), not all research consists of experiments; the two terms are not interchangeable. **Experimentation always involves deliberate manipulation of events**: doing something to someone or something and then determining the effects of what has been done. If, for example, you correlate height with weight to determine the relationship between the two, you have not conducted an experiment because you have not done anything to anyone. If your purpose is to develop a test that will predict stenographic achievement, you are still not experimenting, for you would merely correlate aptitude test scores with achievement test scores. If, on the other hand, you want to know whether a certain mode of instruction changes the relationship between aptitude and achievement, you would be experimenting, for you would have subjected various groups to varying modes of instruction, computed the aptitude-achievement correlations for each of the groups, and tested for significant differences between the resulting correlations.

The foregoing instances illustrate little consensus on a taxonomy of types of research. "Correlational," for example, is sometimes used as a label for a type of research, but a correlation coefficient is a statistic that could arise in various types of research. Prognostic study, to cite another example, is a label based on purposes. In view of the variety of classification schemes for research, one must guard against confusing types of statistics with types of research.

The distinguishing paradigm of experimentation is assessment of the effects **of** something **on** something. As such, only experiments, not any other mode of inquiry, can uncover causal relationships ("If A, then B"). Cancer research illustrates that point. No amount of correlational evidence relating the incidence of smoking to the incidence of cancer can prove that the one causes the other. Proof of cause lies in such tactics as painting the skins of some rats with suspected carcinogenic substances and then comparing the incidence of cancer in those rats with the incidence in genetically comparable rats not so treated. **That** is an experiment. In the same fashion, the merits of various instructional procedures are determined by the measurement of outcomes following varying instruction under carefully controlled conditions, not by surveying the opinions of teachers. Insofar as what goes on in the classroom is a central determinant of educational outcomes, only experimentation can determine what should happen in the classroom.

Terminology of Experimentation. One speaks of variables, a variable being anything that exists in different amounts or forms, such as height, IQ, typing speed, method of instruction, duration of work period, level of anxiety, mode of furnishing knowledge of results, to name just a very few. The experimenter manipulates certain variables and assesses the effects of those manipulations on other variables. The ones the experimenter is free to manipulate are the **independent variables**. Because the outcomes depend on those manipulations, the measures of outcomes are called the **dependent** or **criterion variables.** Furthermore, since the effects of something are always relative to the effects of something else, the independent variable must always be represented by at least two versions of it, called **conditions** or **treatments**. For example, in an inquiry into the effects on typing performance of using a manual versus an electric typewriter, "kind of typewriter" is the independent variable, and "manual" and "electric" are the two treatments or conditions that comprise the independent variable. The dependent variable would be typing proficiency, divisible in this instance into two **criteria** (speed and accuracy). For speed, the **criterion measure** could be, say, words per minute or manuscript pages per hour or time to complete a certain body of work. In turn, some person's **criterion score** could be, say, 43 wpm or 4 manuscript pages per hour or 35 minutes to complete certain work. In some other instance, such as validating an employment test, the validity criterion could be "job success"; the criterion measure, "employer's rating"; and an individual's criterion score, the rating of "excellent." Note the hierarchy of formal terminology from the most general to the most particular: criterion, criterion measure, criterion score. The three terms are **not** interchangeable; for the criterion or dependent variable of "level of anxiety" there are numerous possible measures from among which the investigator chooses one or more as the "criterion measure(s)." For the criterion of typing accuracy, the criterion measure could be number of errors or errors per minute or errors per hundred words or percentage of accuracy, and so on.

Sometimes a particular treatment is contrasted against a **control** treatment. Strictly speaking, a control group is one **not** subject to any deliberate manipulation, of which I know of not a single true instance in business education experiments. Nearly always, business education "control" groups have been subjected to deliberate manipulation (by earlier teachers before the experiment, when not by the investigator during the experiment). Improper reference to experimental and control groups makes it difficult for readers to remember as they read which is the control and which the experimental treatment or group. Labeling the groups according to their treatments, using abbreviations that serve as cues, communicates better and avoids incorrect terminology. Thus, a contrast between distributed and massed practice could refer to D and M groups; a balance sheet versus a journal approach in bookkeeping, to B and J groups; frequent versus infrequent testing, to F and I groups—to give just a few examples.

Other Verbal Considerations. Still another frustration is the use of report titles that do not convey to the reader what the investigation was all about. "Comparison of Shorthand Methods" is an instance exemplifying dozens of business education thesis titles. What methods? The reader scanning a bibliography ought to be able to infer from a title whether the study is of sufficient interest to warrant looking into. At the other extreme, and also to be avoided, are titles of virtually paragraph length that swarm with excess verbiage. "Survey of Later-Life Uses of Personal Typing Skills" is a sufficient title for a study that in fact had a 26-word title. "Stenographic and Clerical Employment Testing Practices Among St. Louis Employers" is better than the 23-word title actually used. "Survey of Typing-Textbook and On-the-Job Production Typing Materials and Development of a Task-Difficulty Scale" is an adequate substitute for the 44-word actual title. Superfluity is evident in such introductory words as: "A study of," "An experiment to determine," "An investigation of." To distinguish surveys from experiments, use "Survey of" in the former; for experiments, "Effects of" or "X versus Y" are among the appropriate introductory words.

Another irritant is "Independent Study" as the only source information in some business education research bibliographies—a flagrant violation of the requirement that a reference include everything the reader needs to know if he or she wishes to gain access to the research report.

Finally, the citation style used in this Bulletin (author's last name and year of publication given in parentheses in the text, with a complete reference list at the end) is the routine style in many professional research journals and books. It is vastly superior to the clutter of footnote citations and the superfluous pedantry of *ibid.*, *op. cit.*, *loc. cit.*, and the like, that make the reader's task more difficult. Footnotes should be reserved for true "asides."

Internal and External Validity as Criteria of Experimentation. The terms "internal and external validity" were introduced by Campbell and Stanley

in their chapter in the *Handbook of Research on Teaching* (1963)—a chapter that became an instant classic and that should be required reading for experimenters. In it, the authors itemize and illustrate the various "threats" to internal and external validity, and they provide a catalog of experimental designs, accompanied by a discussion of the validity of each design.[9] **Internal validity** refers to the confidence one may have that the outcomes result from the manipulated, independent variables, not from other, uncontrolled factors. **External validity** refers to generalizability: the extent to which findings are applicable outside the local experimental setting. Briefly stated, internal validity asks, "Can one trust the findings?"; external validity asks, "Do the findings have general applicability?" Control over irrelevant sources of outcomes and bases for generalizability of outcomes are discussed in more detail later in this Bulletin.

Using the concepts of internal and external validity as a springboard, Campbell and Stanley distinguish between "true" and "quasi" experiments, and they discuss in depth a variety of experimental designs of each kind. The "true" experiment has two defining characteristics: (1) active and deliberate manipulation of all treatments by the investigator and (2) randomization. Lacking either or both of those characteristics, the result is a quasi, not a true, experiment. Some business education experiments are quasi experiments because they lack the first characteristic. An instance is the early research on the Dvorak (Simplified) Keyboard. Outcomes from its use were contrasted with California State Department of Education records of outcomes from standard-keyboard instruction. The investigators did not manipulate, control, or conduct the conventional instruction; they only took criterion scores from records. Whenever one or more of the independent variables has already occurred—not having been subjected to deliberate manipulation by the investigator—we have that class of quasi experiment called **ex post facto**[10] (i.e., after the fact), of which business education contains numerous instances.

All or nearly all educational experiments, including those in business education, have been quasi experiments because they lack the second requirement of the true experiment, randomization at every stage of the sampling process. That unfortunate but unavoidable characteristic has certain consequences for experimental design and data analysis, and it has crucial implications for external validity, which will be discussed later.

---

[9]Most textbooks on educational research since the original Campbell and Stanley publication include updated portions of the original treatment judged pertinent to the audience at which the textbook is aimed—among them, Kerlinger (1973) and Borg and Gall (1971).

[10]The term is not used in an invidious sense. For example, all the experiments on remediation of childhood deprivation are perforce **ex post facto**. Childhood background must be taken as one finds it; the experimenter cannot deliberately manipulate background by depriving some children while enriching others. Another example is inquiries into the causes of and remedies for juvenile delinquency.

Four Prevailing Weaknesses of Business Education Experiments

Four bothersome characteristics are described here in the hope that their frequency will be reduced in future research.

Excessive Length. Some experiments run so long as to prevent attributing their outcomes to the contrasted treatments. One example is inquiries into typewriter keyboard learning, the effects of which are not measured until the end of an entire semester or year of instruction. Between the few weeks devoted to keyboard learning and the terminal testing months later, other, uncontrolled instructional activities swamp the possible effects of keyboard-teaching variations. For another instance, it takes only a few days to assess the effects on typing performance of variations in the amount of warmup. Extended longer than that, dozens of instructional events—far more influential than warmup and varying in uncontrolled fashion—swamp the possible effects of variations in warmup time. It is vital, then, to run studies just long enough to satisfy their purposes, not so long that outcomes could be attributed to events other than the ones under examination. At the other extreme, no reasonable person would suppose that one could assess the effects of differences in the number of new words embedded in shorthand dictation after just a few hours of contrasted dictation practice.

Significant Differences. Although investigators are happy to find significantly different outcomes in studies that hypothesize such differences, it is at times just as useful and important to know that some things make no difference. One little instance is the finding that variations in so-called keyboard approaches (skip-around, horizontal, vertical) do not differentially affect typing performance. Knowing that such variations do not matter permits greater latitude in dealing with other more consequential issues of keyboard-learning methods and materials.

Even so, significantly different outcomes from different treatments is the investigator's usual expectation. To maximize the chances of such outcomes, one must make the variations whose effects are to be examined different enough to have a real chance to make a difference.[11]

Operation of the treatments over a sufficient time period is another factor. Not many experiences have genuine and long-lasting effects if there has been only little exposure to them. Most instructional practices require weeks, even months, to have lasting effects. In the shorthand instance mentioned earlier, contrasting over a period of just a few days many versus few new words in dictation materials would probably have no discernible effects; over a longer time period differential effects would be more prob-

---

[11] In Reha's experiment (see West, 1974, p. 27), which contrasted all of 50-minute periods devoted to production typewriting versus 35 minutes to production typing and 15 minutes to ordinary stroking skills, there were no significant differences in outcomes. Among several possible speculations, perhaps 35 versus 50 minutes is not enough of a difference in instruction to make a difference in outcomes.

able. The point illustrated by the shorthand example does not contradict the earlier comments about excessive length. There, the reference was to the "swamping effects" of more important instructional events; here, the reference is to giving things a chance to take effect.

One-Teacher Experiments. Experiments involving classroom instruction by only one teacher are totally useless if the subject or topic of the research is sensitive to teaching skill or other variations in teachers' behaviors that are not intrinsic to the treatments being contrasted. There are, to be sure, some types of inquiries more or less immune to variations in teacher behavior. Examples are test validation studies and ones devoted to equipment variations and modalities that are essentially teacher-free. Such instances, however, are few and far between. Because most of the important instructional questions are highly sensitive to differences in teacher behaviors, Campbell and Stanley are among many who have insisted that the one-teacher experiment should be forbidden (as research qualifying for a degree or for publication) because it is totally lacking in internal validity. A minimum of two teachers is required, each instructing by all the treatments under examination. Otherwise, one cannot know whether it is the teacher or the treatment that accounts for outcomes. Indeed, the investigator who involves only two teachers is taking great risks.[12]

All too often, the professional journals designed mainly for teachers report one-teacher experiences characterized by high enthusiasm, little if any actual measurement, no awareness of the need to exercise control over possible alternative explanations of the reported outcomes, and no comparison treatment (control or other). The results of such ventures are probably mostly Hawthorne Effect—temporary novelties flowing from temporary enthusiasms.

While adventurous teachers, willing and eager to try new things, are greatly to be encouraged, it is neither helpful nor accurate to call their activities "research." Trained technical skill and knowledge are required. In so-called "action research" readers may be wrongly led to suppose that individual experience is a basis for sound generalizations and that ordinary classroom teaching experience provides accurate insights into the enormous

---

[12] A sample of one has no variance and, for that reason, is not susceptible to statistical analysis. The investigator who uses a sample of only two (teachers) is statistically naive. If there should be a nontrivial difference in the skills of the two teachers, analysis will result in significant "interaction" (between teachers and treatments), thereby reducing or even precluding the chances of significant treatment effects. With more teachers involved, differences in teaching skill will tend to balance out, leaving differences in outcomes, if any, more likely to be attributable to the treatments. In all research "there is safety in numbers." In examining research reports of the kind discussed here, look especially for a test of statistical significance of "teacher-by-treatment interaction."

subtleties of learning processes. The proper meaning of "action research" refers to local inquiries whose outcomes may not have and are not intended to have any generalizability beyond the local situation. In all other respects the requirements for proper design and conduct of the inquiry and for appropriate data analysis apply equally to the action researcher and to the most experienced specialist.

Prediction of Outcomes in Advance. Sir Isaac Newton once remarked, in explaining his towering achievements, that he stood on the shoulders of geniuses—the scientists and thinkers who preceded him. Educational research also stems from earlier investigations in various fields, rarely charting new paths, but instead adding a few inches to a path already miles long. Indeed, if there were not something of an iron curtain between teacher education and the basic research of the experimental psychology of learning, the applied research in education would be more productive. For example, familiarity with the generalizations arising from the immense body of psychological research on distribution of practice would permit the design of inquiries into the scheduling of various topical components of given business subjects. So designed, such inquiries would have a real chance of payoff, and the general direction of their outcomes could be more or less predicted in advance. The experimenter who cannot anticipate the general direction of outcomes is not very sophisticated. This is not to say that surprises never happen or that there are no ventures into uncharted regions. Rather, most educational research merely adds refinements or greater precision to general phenomena that are already well known. With occasional exceptions, business education research has been derivative, not original.

The pertinent inferences are: (1) You have to know the past in order to attempt sensible research; (2) Often, the general direction of outcomes should be predictable in advance; (3) Avoid inquiries whose outcomes are so certain from past research as to make them a waste of time. For example, reinforcement is such a solidly established requirement for all of learning that it is pure waste to look into the effects of reinforced versus nonreinforced responses in any business education subject; the answer is totally foreordained. There are instances, however, in which some outcomes are predictable while others are not. One example is an inquiry that has a bearing on sight versus touch typing among beginners (West, 1969a, Chap. 4). Psychological studies going back many years had already shown that, regardless of the motor skill involved, beginners cannot rely on their muscular sensations as indices of correct motions. The inevitable outcome of the typing study conducted decades later was that beginners could rarely tell, purely from the sensations in their finger muscles when deprived of vision, when they had made a stroking error. Thus, strict insistence on touch typing at the start is seriously wrong because it robs the learner of reinforcement, thereby violating the first and foremost rule for all of learning. What was not predictable in that same typing study (because no earlier research had raised the question) was the rate of increased dependability of muscular sensations as skill increases or, to mention a particular finding, that even

100-wpm typists can depend solely on their muscular sensations only about half the time.

Major Features of Experimentation

Considered here are: (1) Sampling and randomization, (2) Wording of hypotheses, (3) Controlling for irrelevant sources of variation, (4) Valid and reliable criterion measures, and (5) Other statistical requirements.

Sampling and Randomization. Random selection of participants at each stage of the sampling process is one of the two hallmarks of the "true" experiment, but it can rarely be accomplished in educational experiments. Nearly always, one deals with what is at hand: particular schools, particular teachers, particular classes, particular students. We use willing or volunteer schools and teachers; and volunteers, as all authorities point out, are not a representative sample of all schools, teachers, and students. The theoretical ideal is to select school systems at random from all school systems, schools at random from the selected systems, teachers at random from the selected schools, classes at random from those of the selected teachers, and, finally, to assign students at random to the selected classes. Only such a sequential process provides high confidence in external validity, in the applicability of one's findings to students, teachers, schools, and school systems in general. Only enterprises planned and funded on a national scale have been able to use such impeccable procedures. An example is *Project Talent* (Flanagan et al., 1962), and it was a survey, not an experiment.

In the absence of randomization at every stage of the sampling process and faced with the question (in some hypothetical experiment) of whether the 176 advanced shorthand students in the 6 classes of a certain 3 teachers are representative of all advanced shorthand teachers and students, evidence for external validity, for generalizability beyond the local participants, must necessarily take another route: the route of replication, discussed later.

Even within the limitations of educational experimentation, randomization at the final stage is possible and must be used. Given two treatments and a pair of classes for each of several teachers, you can literally toss a coin to determine **which** class of **which** teacher receives **which** treatment. If classes are already formed, so that the experimenter is not free to assign students at random to classes, you have "cluster sampling": The unit of sampling is the class, an "intact group," whereas statistical analysis is predicated on "element [student] sampling." The implications for statistical analysis are discussed later.

Among the several stages of sampling, the compelling requirement is for random assignment of students to classes (treatments). The usual assurance of school officials that students are assigned to various sections of a course at random requires checking; explicit and pointed inquiry is needed to pro-

tect against instances of homogeneous grouping not recognized as such.[13] Such instances should be discarded and substitutes sought.

The uncertain representativeness of volunteer participants and of intact groups often and properly leads to what is called "differential analysis": examination of the pertinent characteristics of students in intact groups and to analysis of data for subgroups stratified after the fact according to the selected characteristics. In surveys, stratification is commonly done in advance and is for the purpose mentioned earlier. In experiments, stratification is done after the fact, usually at the stage of data analysis, and is employed when it is suspected that some effect may be present among some students but not among others. For example, Method A might be better than Method B for bright students, but the reverse might obtain among less able students.

Beware, however, of **regression effects**: the tendency of those at the extremes to move toward the mean—extreme scores on pretests tending to go with less extreme scores on posttests, as Borg and Gall explain and illustrate (1971, p. 408). Couple that universal phenomenon with the negative acceleration of the acquisition curve for skills (gains being made rapidly at the start, more slowly later on), and it is easy to see that poor students will generally gain more than good students (whatever the instructional method) because gains are easier to make at low levels of skill and knowledge. Touting some method of instruction as effective for poor students runs afoul of that "more room for gains" fact.

On the other hand, a treatment reported as ineffective among able students might merely reflect **ceiling effects**: little room for further gains. Further gains may be limited either because those already at high levels of skill need more time to make gains than the time span of the inquiry permits or because pretest scores are already close to the maximum possible score (as with a pretest score in the 90's when pre- and posttests have a maximum possible score of 100). For these and other technical reasons, avoid "gain (or change[14]) scores" as criterion measures. Instead, pretest participants and use analysis of covariance, as explained later in this Bulletin.

If, however, data for subgroups are analyzed, avoid "so what?" or "tag along" bases for stratification: ones that, no matter what the outcomes, have no practical implications. It is more than doubtful, for example, that different outcomes for boys and girls will lead the schools to homogeneous

---

[13] Examples are: stenographic majors assigned to the 10 a.m. typing class, but clerical majors to the 11 a.m. class; morning session students (the higher grades) to the 1 p.m. class, but afternoon-session (lst- and 2nd-year) students to the 2 p.m. class; bookkeeping majors to one office practice class, but secretarial majors to another.

[14] "Change score" is the preferred term because "gain" connotes a higher or larger number as the measure of performance. However, if the number of errors is a criterion measure, improvement or gain in accuracy is represented by a smaller number than originally (e.g., a reduction from 6 to 3 errors).

grouping by sex or to different instructional methods for the two sexes. Analysis by sex is appropriate and important only in basic research on theoretical issues specifically addressed to sex-based differences (whether biological or cultural). In such instances, that variable is discussed as part of the rationale of the experiment, not dragged in as an afterthought, in irrelevant instances, merely to fatten up the body of results.

In summary, note that in surveys stratification is done in advance, whereas in those experiments in which it is neither possible nor desirable to group participants in advance on the basis of selected background characteristics, it is done at the end—if and when important differences in background characteristics of persons in intact groups have been identified.

Wording of Hypotheses. In the context of the present discussion, a hypothesis is a statement of anticipated outcomes. An "experimental hypothesis" would be worded to convey actual expectations: for example, that Method A will be superior to Method B. For convenience in statistical analysis, however, it is conventional to state a "statistical hypothesis," one worded in "null" form stating, for example, that there will be no difference between the outcomes of the treatments. Whatever the form, the wording of hypotheses should conform to the intended criterion measure. A negative instance is a study that hypothesized that Method A will "sooner" bring students to a certain level of proficiency than Method B. With such a hypothesis one expects that a certain level of proficiency will be stated and that "soonness" or the time to reach that level will be the criterion measure. What was actually compared was the difference in proficiency after fixed amounts of training. No harm was done in that experiment because the real question is answered as well in one form as the other. Nonetheless, the requirement is that hypotheses be worded in keeping with one's criterion measure(s).

Controlling for Irrelevant Sources of Variation. Laboratory experiments tend to have high internal validity; each feature, large or small, is rigidly controlled, and potentially contaminating factors are carefully excluded. The classroom is decidedly less docile, less manageable than the laboratory. Managing matters so that the experimenter can rightfully attribute outcomes to the treatments that make up the independent variable(s)—unconfounded, uncontaminated by other events that could influence the results—is a difficult task. But it must be done well if findings are to be taken seriously. Control over irrelevant sources of variation is the heart of experimentation.

There are three types of control mentioned in all the literature on the conduct of experiments: (1) randomization, (2) experimental control, and (3) statistical control. A subclass of experimental control frequently enough violated to deserve special mention is (4) monitoring.

1. **Randomization** is operating when you toss coins or consult a table of random numbers to select participants and to assign treatments to them. It also applies to events that could affect outcomes, but which can reasonably be expected to have equal effects on all treatments and can therefore be

ignored. Absenteeism is a good example of many such factors. Unless one of the treatments is so painful as to merit the famous drama review reporting that "The audience stayed away in droves," one may reasonably assume that absenteeism will be at about the same level for all treatments—requiring no after-the-fact examination.[15] In general, any event that could affect results, but that is likely to be equally present in all treatments, may be said to be randomized and, for that reason, need not be taken into account.

2. **Experimental control** was illustrated earlier by designs employing several teachers, each employing all treatments (so that it is possible to measure the portion of outcomes attributable to differences in teaching skill), and by experiments that run just long enough to satisfy their purposes. Then there are those powerful designs in which each subject serves as his or her own control, as in some of the manual versus electric typewriter experiments. In them, all students are taught both machines, variously in E-M or M-E order [providing **not** C(ontrol) and E(xperimental) groups, but E-M and M-E groups]. Another instance is inquiry into the effects of four different amounts of warmup at the typewriter, all students being subjected to all amounts in counterbalanced order, a different amount each day on each of four successive days.

Counterbalancing is a standard aspect of experimental design that equalizes practice, fatigue, and transfer effects. One example is a pair of timings in a typing class administered in a-b order to one random half of the class, in b-a order to the other random half of the class. A more general use of counterbalancing applies when all treatments are to be administered to all subjects. In such instances the order of administration varies, while the sequence within order is usually the same. For example, with three treatments (e.g., 1-, 2-, and 3-minute rest periods between work periods), 1-2-3, 2-3-1, and 3-1-2 orders would be assigned, respectively, to randomly selected thirds of the participants. If sequence within order is suspected as likely to have irrelevant effects, three additional sequences (1-3-2, 2-1-3, 3-2-1) would be used, and participants would be randomly divided into sixths. When the events are too numerous to make practicable the use of all possible sequences (4 events can be sequenced in 24 ways), some of the various ways are used—as in a 4-item test booklet whose subtests are assembled in 8 of the 24 possible sequences. The latter instance is also a protection

---

[15] Absenteeism, however, is a major index of motivation. If the contrasted treatments are expected to affect motivation differentially, it should **not** be assumed that absenteeism is approximately equal in all groups. Instead, careful attendance records should be kept, and attendance data should be analyzed as one of the dependent or criterion variables. Of course, time of day also affects school attendance, which tends to be somewhat poorer for the earliest morning and latest afternoon classes—a tendency that needs to be taken into account.

against cheating. Booklets so sequenced can be distributed to insure that no person is working on the same test item at the same time as his or her immediate neighbors. The rule for experimenters is: **When counterbalancing is desirable, it is mandatory.**

Once a design appropriate to one's purposes has been selected, the major requirement is for control over factors that interact directly with the independent variable(s), ones that must be held constant. It should be obvious that an inquiry into "kind of practice" (comparing different practice materials) must hold constant "amount of practice" on each of the various kinds of materials. Similarly, if amount of practice is the independent variable, kind of practice must be held constant. In the same fashion, you cannot test motivational tactic A plus amount of instruction X against motivational tactic B plus amount of instruction Y. At the end, you could not know what accounts for the outcomes, whether the differences in motivational tactics or in the amount of instruction. There are no cookbook rules for guarding against such impermissible confounding of variables. Ordinary common sense, plus a knowledge of what factors tend to influence or interact with others, is required.

There are, of course, "factorial" designs that are used precisely because interaction effects are expected. Much agricultural research, for example, examines the effects on plant growth of different fertilizers, used in different amounts, in different soils—all fertilizers being used in all amounts in all soils. In education and psychology, distribution-of-practice experiments are often so designed, various combinations of duration of work period and rest period between work periods being employed.

3. **Statistical control** can take either of two forms. One is partial correlation, used to examine certain relationships while holding other factors statistically constant. An example is measures of relationships between theory and transcription scores in shorthand, which "partial out" the effects of differences in IQ (regrettably not done by the various investigators of theory/transcription issues). The better tactic—immensely powerful and widely applicable—is the analysis of covariance to account for pre-experimental differences among learners. For example, differences in terminal achievement in intermediate shorthand would be analyzed in relation to differences in shorthand pretest scores upon entering intermediate shorthand.

With rare exceptions, covariance analysis is far preferable to matching students, whether in advance or afterwards. Among several serious objections to matching, one is that it is often impossible to find matches at the extremes (the very high and low scores); matching thereby loses cases and leads to a sample that does not represent a population that indeed includes extreme cases. Second, it is virtually impossible to match persons on more than one or two features, and it is axiomatic that most complex phenomena rest on more than one or two pre-existing conditions. Thus, matching fails to represent all the pertinent variables in the population, resulting in a less, rather than a more, representative sample.

A mild case can be made for matching when N (the number of scores, persons, events, etc.) is very small and when only small differences in outcomes are expected. In education, however, small differences rarely have practical consequences, so matching is for that reason counterindicated. If you must match, rank students in order and assign them in alternation to treatments, thereby losing no cases. Then, be certain to use the mode of statistical analysis that applies to matched groups or individuals; otherwise, what is called a Type II error could result: failure to reject a null hypothesis when there is in fact a significant difference between treatments.

Finally, the basis for matching and the pre-measure in covariance analysis must meet two criteria: satisfactory reliability and a nontrivial correlation with the criterion measure. The investigator in a production typing experiment who matched students on the number of straight copy errors failed to meet both criteria: Measures of straight copy errors have notoriously low reliability and not the slightest relationship to production typing proficiency. Similarly, because IQ has little relationship to ordinary copying skills, analysis of straight copy scores in relation to differences in IQ is pointless. In general, there are no background factors so powerful as measures of initial status as task performance measures themselves.[16] Avoid pre-measures that have uselessly low relationships to criterion performance.

4. Monitoring (an aspect of experimental control) refers to the investigator's responsibility to insure that what is supposed to happen in the classroom indeed happens and that what is not supposed to happen does not happen. One can specify what participating teachers are and are not to do, but one cannot assume that they will act according to specifications, without exception. When engaged in highly novel activities, they may sometimes slide, without awareness, into familiar ways; or they may second-guess the investigator and make "improvements" that have the effect of contaminating the treatment. Indeed, Murphy's Law ("If anything can go wrong, it will.") has notable applicability to classroom experimentation.

Specifications for teachers involved in experiments must be spelled out in painstaking detail; one must anticipate and provide for every eventuality. When novel procedures are to be employed, it is necessary to pretrain the teachers in the use of those procedures. Even then, teachers' activities have to be monitored periodically by first-hand observation. Sometimes, inquiries of students can serve as a check. Primarily, it is often possible to tell from students' daily work whether teachers and students have acted according to specifications. If some shorthand experiment requires, say, three repetitive dictations of the same copy on certain days, students' notes will

---

[16] For example, straight copy scores on the first day of class, before instruction (students use any modes of operating the typewriter whatever) could easily be better predictors of end-of-semester performance than are background factors such as age, sex, IQ, and the like.

show whether there were three dictations. The investigator should therefore collect and examine student work, especially at the start, in order to nip in the bud any possible student or teacher misunderstanding.

Few reports of classroom experiments specify the details of the close monitoring that is required. The susceptibility of the classroom to contaminating factors often accounts for the frequency of "no-difference" outcomes. Contrasts derived from some of the most fundamental findings about learning processes (established under the rigorous controls of psychological research in laboratory-like settings) often fail to meet expectations precisely because the classroom "waters" are so frequently "muddy." A large part of that muddiness could be strained out under tight monitoring applied to explicit specification of permissible and impermissible classroom activities.

Valid and Reliable Criterion Measures. No amount of careful operation of independent variables can achieve one's purposes unless the dependent or criterion variable(s) used to assess the outcomes are valid and reliable, unless they in fact measure the outcomes pertinent to the hypothesis under test and unless they do so with reasonable precision. Low validity is evident in those production typing studies whose criterion tasks used prearranged materials, rather than ones requiring the examinee to make decisions about the placement or arrangement of materials on the page. Low reliability results from tests that are too short to provide even moderately stable measures of stroking accuracy at the typewriter. Irrelevance to the treatments (invalidity) is apparent in investigations of the role of shorthand theory in stenographic proficiency that penalize for spelling and punctuation errors in typewritten transcripts of business-letter dictation. Because shorthand theory has no bearing on language errors, the latter should not be used as criterion measures.

The requirement, in short, is to confine criterion measures to those that could be differentially affected by the treatments that make up the independent variable. Nearly all dependent variables or criteria in experimental research are measurements, whose credibility as indices of the effects of independent variables rests entirely on their validity and reliability. On the validity and reliability of one's criterion measures, the experimenter is not permitted to be silent.[17]

A Miscellany of Other Statistical Aspects. Discussed next are: (1) Statistical and practical significance, (2) Interpretation of correlational data, and (3) Sample size (especially in relation to the use of intact groups).

---

[17] See West (1975) for the meaning of validity and reliability as applied to tests and for details on the merits of various testing procedures for the measurement of typewriting and stenographic proficiency.

1. **Statistical and practical significance** are the payoff indices of experimentation because unanalyzed data are useless. There are "chance" effects and there are "real" effects, and only analysis can separate the two.[18] The appropriate mode of analysis can vary with variations in experimental design. The wise experimenter therefore consults a statistician **in advance** to make sure that the design is one that will satisfy the purposes of the research and that the resulting data will be susceptible to appropriate analysis.

Furthermore, it is important to recognize that statistical significance is a precondition for, but by no means a guarantee of, practical significance. A "chance" outcome is one that is well within the variations that characterize random events. For example, although the theoretical expectation is for five heads and five tails in ten tosses of a coin, an outcome of six heads and four tails would surprise no one and would not justify the inference that the coin is biased. A "practically significant" outcome, on the other hand, is one large enough to do something about, to take action on, rather than one that engenders an "OK, but who cares?" reaction. For example, a correlation coefficient of .06 has less than 1 chance in 20 of being due to "chance" if it is based on 1,000 cases; but it is so close to zero as to have no practical significance. The slang expression "There oughta be a law against..." applies to investigators who report only the probability (i.e., significance) levels of their findings without specifying the findings themselves, whether correlations, means, differences between means, or whatever. In any event, to say that statistical significance is a precondition for practical significance means that without statistical significance, the question of practical significance may not be raised. Without statistical significance there can be no practical significance. A statistically significant outcome, however, might or might not be judged to have practical significance.

2. **Correlation coefficients** are measures of concomitance, not causation. The appreciable correlation between body weight and body height certainly does not mean that weight causes height. Nor do high correlations between shorthand theory and shorthand transcription scores necessarily imply a direct (causal) relationship between the two. More likely than not, both are related to a third factor, general intelligence, which would have to be "partialed out" if an unconfounded measure of theory/transcription relationship is to be supplied. In any event, the suspicion of causation sometimes permitted by high correlations has to be tested by experimentation, which is the **only** mode of inquiry that can assess cause-and-effect relationships (as illustrated by the cancer studies mentioned earlier).

3. **Sample size** relates to (a) statistical significance and the precision of estimates of population parameters; (b) certain experimental designs; and,

---

[18] Kerlinger (1973) provides exceptionally clear and simple illustrations of the major statistical measures applicable to the testing of hypotheses as part of his comprehensive treatment of educational and psychological research.

crucially, to (c) mode of data analysis for the intact groups (pre-formed classes) used in practically all classroom experimentation.

a. **Statistically significant values** in small samples are exactly as real as those in large samples, because all statistical tests take sample size into account. (In fact, small-sample significance can be impressive just because it is difficult to achieve significant effects in small samples.) Increases in sample size, however, increase the reliability or precision of estimates of population parameters. Measure 10 men and 10 women (in order to estimate the height difference in the population of all men and women), and the estimate for the difference in height might be 5 inches ± 2 inches. Measure 100 men and 100 women, and an outcome such as 5 inches ± ½ inch might result. The small sample estimates the difference as between 3 and 7 inches; the larger one, as between 4½ and 5½ inches. Both samples provide equal assurance that heights differ significantly, but the large sample supplies a more accurate estimate of the size of the difference.[19]

In contrast to the function of sample size in surveys and in the light of the vital distinction between statistical and practical significance, Tatsuoka (1969, p. 480) is one among many who inveigh against too-large samples in experiments, as follows:

> The uninitiated may...[think] that (within budgetary limits) the larger the sample the better.... But in experimental studies [in contrast to surveys]...using too large a sample can actually be deleterious. This is true because, given a large-enough sample, a relationship of the minutest degree may show up as "statistically significant" which is at best a nuisance and at worst can be actually misleading. It is therefore desirable to determine an appropriate sample size which is large enough to detect (with a preassigned probability) a relationship of practically important magnitude and is yet small enough to preclude from "significance" a relationship of trivial degree.[20]

b. **Certain experimental designs** require equal Ns, equal numbers of persons or events in each group under study. In such instances, the procedures do not require starting with equal Ns. Instead, when all the data are in hand, one randomly discards cases from the larger groups in order to make all Ns equal to that in the smallest group. But from the outset you should make the smallest group large enough because (i) there will almost certainly be attrition (loss of cases through absenteeism, missing data, etc.)

---

[19]The estimate would of course be accompanied by a probability statement; e.g., that the chances are 95 in 100 that the mean difference in height lies between 4½ and 5½ inches. Such a statement illustrates the meaning of the last sentence of Footnote 3.

[20]Tatsuoka (p. 481) cites a textbook by Hays that describes how to determine optimum sample size.

and because (ii) the dice of statistical significance are loaded against small samples (fewer than 30). Whether or not Ns need be equal, one should make every effort to end with at least 30 representing each treatment.

c. **Intact groups** (pre-formed classes to which the investigator has **not** assigned students at random) are the bane of classroom experimentation. In such instances, the "element" for statistical analysis is not the student but the class. Unless there is very good assurance that the school has indeed assigned students to classes at random rather than on some homogeneous basis, a minimum assignment, **at random**, of at least ten classes to **each** treatment has been recommended (Tatsuoka, 1969, p. 480). The units of observation then properly become the class means, **not**, improperly, individual student scores. Obeying Tatsuoka's recommendation, however, would wipe out most classroom experimentation, because access to and management of at least ten classes per treatment is almost always beyond the time, talents, and funds of the doctoral candidates who produce so large a proportion of all business education research. In solving the impasse created by cluster rather than element sampling, one turns to replication as the available if not perfect means of improving external validity.

Replication as a Basis for External Validity

Replication, the independent repetition of an experiment, has been called "the cornerstone of scientific inquiry" (by Bauernfeind, as quoted in Borg and Gall, 1971, p. 290). In the sciences replication applies to even the smallest details. Any important new breakthrough is quickly checked on by other scientists, who repeat the original experiment with great exactness (a process that demonstrates the necessity of a detailed account of procedures). In educational circles, however, exact replication is rare and, among doctoral candidates, forbidden (except in periodic trend studies or in research reviews, which are not experiments). The doctoral candidate is required to do something new or, at very least, to engage in some nontrivial variation on some earlier-investigated theme or to correct earlier deficiencies in design or procedures.

Lykken, quoted by Borg and Gall (1971, pp. 290-291), distinguishes among three types of replication: literal, operational, and constructive. Literal or exact replication is forbidden the doctoral candidate. Operational replication duplicates only the sampling and experimental procedures of earlier research and tests whether the outcomes duplicate those of earlier work. Constructive replication gives the replicator a free hand to test by his or her own sampling, experimental, and measurement procedures findings reported in earlier research. The prevailing tendency of business education research to provide different studies of the same general topic illustrates constructive replication.

That tendency permits making a virtue of necessity. As different experimenters involve different students, teachers, and schools and employ

greater or lesser variations in independent and dependent variables associated with some one broad question, one has increasing confidence in the external validity or generalizability of outcomes whenever they tend to be in the same direction (if not of the same size). In such instances, generalizability applies to the broad issue variously implemented in the constructive replications, not necessarily to their fine details. Also, only external, not internal, validity is established by replications that have consistent outcomes. Twenty badly designed and conducted studies with comparable findings have no internal validity and for that reason can prove nothing. In any event, business education research contains numerous instances of consistent outcomes from constructive replication.[21]

From the replications-with-variations that characterize much educational experimentation, it should be apparent that major generalizations are established by accretion, each inquiry adding another brick to what can then become a solid structure, a pervasive generalization applicable to teachers and learners in general. Such a recognition should be reflected in the reporting of research, as should new findings that are startlingly out of keeping with the bulk of earlier findings.

Pitfalls in Research. A finding strikingly different from that of other studies of the same phenomenon requires especially close examination. Perhaps it arises from what Barber in his chapter on "nine pitfalls in research" (1973) calls "Investigator Fudging Effect," a euphemism covering everything from woeful ignorance to outright fraud. The intricacies of statistical anaylsis of data are one common trap, especially among investigators whose statistical understandings are limited and who turn over data analysis to a computer center, whose staff has no responsibility for interpreting the outcomes of analysis in relation to experimental hypotheses. In any

---

[21] Among several in typewriting (see West, 1974, for details)—all involving from many hundreds to many thousands of students, correspondingly large numbers of teachers, classes, and schools, and numerous internal variations in procedures—one is the uniform finding of the ineffectiveness of nonsense drill for keyboard learning. Other uniform findings show the total uselessness of specialized drill materials for promoting stroking accuracy and that the number and kinds of stroking errors are independent of the characteristics of the materials. The latter finding necessarily follows from the generalization established by psychological research, which states that in all perceptual-motor skills accuracy depends on controlling the response rate, on working at the right speed—not on the characteristics of the work or practice materials. Another finding—of all the several studies except Reha's (see Footnote 11)—is that vastly superior production typing proficiency (as well as equal straight copy proficiency) follows from an early and nearly exclusive focus on production tasks, not on ordinary stroking skills. Comparable unanimity of findings is also found in surveys. All show the low relevance of the conventional high school bookkeeping curriculum to the actual job duties of beginning bookkeepers, especially under computerization of accounting records (West, 1973). Another uniform survey finding is of the prevalence of longhand and mixed-type-and-longhand as input materials among employed typists (West, 1974).

event, novel findings demand explicit consideration by the investigator,[22] who, in reviewing the earlier literature (extensively in a dissertation, more briefly in a published article) and his or her own findings in relation to earlier findings, has to bring to bear sophisticated, not superficial, analytical and critical faculties.

Summary of Major Experimental Requirements

1. There must always be at least one alternative treatment against which the effects of the given treatment are compared. Single-treatment investigations are useless. Do not label the alternative treatment a "control" treatment unless it involves no manipulation of events.

2. To maximize the chances of statistically significant differences in outcomes, make the treatments sufficiently different and maintain their operation long enough to give them a chance to have differential effects. On the other hand, investigations of independent variables whose effects on outcomes are likely to be modest in relation to the effects of other, uncontrolled instructional events should not run so long that their possible effects are swamped by the effects of the more important events.

3. Although random sampling of participants at each stage is rarely possible in experimentation, at a minimum (a) assign treatments at random to the classes of the participating teachers, and (b) if you cannot assign students at random to treatments (classes), check carefully with school officials to determine whether such assignment has indeed been on an essentially random basis.

---

[22] An instance of this pitfall is a purported finding of near-zero correlations between IQ scores on the one hand and, on the other, shorthand theory and transcription scores—side by side with near-perfect correlations between theory and transcription scores. Such outcomes fly in the face of the long-established phenomena of high relationships between verbal tasks and measured IQ and the rarity of high correlations between a complex task and a component of that task. Specific to stenographic skills, all the numerous earlier studies routinely showed appreciable correlations between IQ and various measures of stenographic proficiency and, as well, the prevalence of correct transcript from wrong shorthand (West, 1969b). The failure of the investigator to discuss the novel findings in relation to the diametrically opposite body of earlier findings illustrates one kind of "Investigator Fudging Effect" and requires rejection of the novel findings and of the inferences for stenographic instruction drawn from those findings.

Concerning research on stenographic learning, an interesting question awaiting an answer is whether the same outlines are written correctly in both theory tests and under the press of continuous-matter dictation. In designing such studies it would be important to distinguish between shorthand "theory" and shorthand "vocabulary." The former properly refers to the application of a rule of the shorthand system to the writing of an outline—an ability that can be tested only by presenting the examinee with words never before read or written in shorthand. To write correctly outlines that have been encountered in earlier learning merely measures memory for earlier-practiced materials. Actually, all the inquiries into so-called theory mastery have used vocabulary, not theory, tests. If shorthand "vocabulary" is indeed the proper focus of interest, the word "theory" should be discarded; reserve it for measures of the ability to apply the rules of the shorthand system to new outlines.

4. The first and foremost requirement is for internal validity, for assurance that outcomes are indeed the result of the manipulated variables, not of other possible causes. Control over such "other" possible causes of outcomes takes three forms: randomization, experimental control, statistical control. The most obvious instance of experimental control (over variations in teaching skill) is the requirement for at least two (but preferably more) teachers, each using all treatments—each treatment being applied to different students or, under counterbalanced designs, all treatments being applied to all students. As an instance of statistical control, matching of students is rarely desirable. Analysis of covariance examines final differences in relation to initial differences and is far superior. In any event, the covariate (measure of initial status) or the basis for matching must have both high reliability and a significant relationship to the criterion measure(s). The most powerful pre-measures are task performance measures (**not** such background factors as sex, age, IQ, and the like), preferably ones like the criterion (final) measures, if at all possible in the light of the question under investigation.

5. All procedures must be described in sufficient detail to permit an independent replication or repetition of the experiment. Vague mention of a contrast with a "traditional" or "control" method will not suffice. Moreover, whatever the procedures, classroom experiments require explicit monitoring to insure that specified events take place and that forbidden events do not occur.

6. Unless there is very good assurance of random assignment of participants (students) to treatments, either by the investigator or by school officials, the resulting "intact groups" require the use of class means, not individual student scores, as the elements for data analysis. With nonrandom (homogeneous) assignment of students to intact groups, an N of no fewer than ten classes per treatment is recommended.

7. For treatments that might have differential effects on students with different characteristics, avoid "so what?" bases for separate analyses—ones that have no practical implications (e.g., sex). In examining possible differential effects on learners at different levels of ability, be cognizant of "regression effects" and of "ceiling effects" for able students, as well as of the truism that the poorest students have the most room for gains. Avoid "gain" or "change" scores as criterion measures.

8. Statistical significance is merely a precondition for considering practical significance: whether an outcome is of sufficient absolute size to be worth doing something about. Given statistical significance, silence on practical significance is impermissible. That caveat is particularly important with large Ns because, with large Ns, even minute differences can be statistically significant.

9. Beware of such prevalent statistical misunderstandings as: (a) inferring causation from concomitance (i.e., from correlations) or assuming a real relationship between x and y when, in fact, both are related to z; (b) equat-

ing practical with statistical significance or failing to report the actual "obtained" values (means, correlations, and the like), whatever their significance levels; (c) rejecting small-sample findings for reasons of sample size (all statistical analysis takes sample size into account, and larger samples merely provide more precise measures). Principally, consult a statistician in advance and refrain from any involvement in experimentation whose sample size is so small as to make significant outcomes highly unlikely.

Requirements for Research Reporting

Formal inferences and recommendations must be confined to those supported by one's findings, considered together with those of comparable earlier studies. Inferences and recommendations going beyond one's findings are, unfortunately, not uncommon and are to be guarded against.[23] When speculation beyond one's findings is indulged in, it should be clearly labeled as such and should not extend to wild flights of the imagination unanchored to fact or theory. Mark Twain's amusing remark about excessive speculation (in his *Life on the Mississippi*) is: "One gets such wholesale returns of conjecture out of such a trifling investment of fact."

The treatment of one's own work as the final and definitive word on a given question is one hallmark of the novice. Knowledge grows by accretion. Because most educational research tends to play variations on general themes represented in earlier research, it is important to discuss, late in the report, the implications of one's own findings in relation to those of earlier research. That requirement is especially important because most educational experiments are quasi, rather than true, experiments. As such, evidence for external validity, for generalizability beyond the local setting, comes from a body of studies of which one's own work is merely one part. Only when most studies on some topic—despite variations in internal details—have outcomes in the same direction is the generalization that is the objective of all inquiry attained. Discussion of one's own findings in relation to those of others often permits identifying those differences in pro-

[23] One instance is a study that supplied neither more nor less than a rank order of words according to frequency and whose investigator recommended that the high frequency words be subjected to heavy practice in typing and stenographic instruction. There is, however, nothing about a word list *per se* that has anything to do with its possible uses. That issue is decided on other grounds and, as it happens, the pertinent evidence is strongly against inordinate focus on a small vocabulary of common words (West, 1969a, pp. 153-163). In another study, the investigator reported an average of 2.3 errors per minute among typists and a few pages later recommended a D grade for those who make more than 2 epm—thus guaranteeing a grade of D or less to more than half of all typists. In the same study various types of stroking sequences were purportedly found to affect stroking errors, and drill materials based on stroking sequences were implicitly recommended for accuracy improvement. First, whether or not accuracy varies with stroking sequences has nothing to do with whether stroking-sequence practice will improve accuracy. Second, as it happens, the dozens of just such experiments showed letter-sequence drills (and everything else by way of specialized materials) to have not the faintest beneficial effect on stroking accuracy (West, 1969a, pp. 257-269; or 1974, pp. 14-15).

cedural details that might account for differences in outcomes or, equally important, establishing a phenomenon so rugged that it is impervious to variations in details.

Finally, a crucial reporting requirement is for explicit details on all procedures in all treatments, not merely the "experimental" treatment. For one thing, conflicting outcomes in studies presumably dealing with the same phenomenon are nearly always attributable to procedural differences. The reporter owes the reader a basis for identifying what might account for differences in outcomes. Additionally, verification of findings through independent replication, with or without variations, is a leading canon of science, but it cannot be done unless procedures are described in sufficient detail.

Impact of Research on Practice

Krathwohl, in discussing the "perceived ineffectiveness of educational research" (1974), contrasts the personalities of those educators attracted by the scientific model with those humanistically oriented personalities drawn to education in order to serve human needs. In the same vein, C. P. Snow deplored the lack of communication between the scientists and the humanists in his phrase the "two cultures." And at the turn of the century William James divided mankind into the tough-minded and the tender-minded. All three gentlemen were saying the same thing; but in educational affairs the effect has been devastating. It has led to masses of teachers not only ignorant of the proper bases for educational behaviors, but hostile to objective evidence as a basis for instruction and a prey to the delusion that personal teaching experience is a basis for generalization. Moreover, as many critics have observed (e.g., Smith and Meux, 1964; Travers, 1964), teacher education has dealt mainly with peripheral notions rather than with central concepts about learning and is "replete with useless residues" (Smith and Meux, 1964, p. 108).

Although no one who is familiar with the evidence on the development and modification of personality would propose to try to change the personalities of those drawn to teaching, the fault lies in the teacher education process itself. No improvement can be expected so long as undergraduate teacher education tends to consist of take-on-faith prescriptions that play back existing practices, with challenge to existing notions reserved for graduate education (and, indeed, the higher levels of graduate education). Unless research findings are brought to bear on the education of teachers from the outset—findings that come to grips with the teacher's central responsibility for conducting instruction—we shall continue to cling to ancient superstitions and to make desirable changes at a snail's pace, rather than with dispatch. There is nothing democratic about evidence; objective data on instructional processes are not subject to acceptance or rejection by majority vote based on personal experience, either as teacher or student. The plea here is for the infusion of reputable evidence into teacher education from the earliest undergraduate years. To justify that infusion, the research has to have high internal and external validity.

References

Babbie, E. R. *Survey research methods*. Belmont, Calif.: Wadsworth, 1973.

Barber, T. X. Pitfalls in research: Nine investigator and experimenter effects. In R. M. W. Travers (Ed.), *Second handbook of research on teaching*. Chicago: Rand McNally, 1973.

Borg, W. R. and Gall, M. D. *Educational research* (2nd ed.). New York: McKay, 1971.

Campbell, D. T. and Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

Flanagan, J. C., Dailey, J. T., Shaycroft, M. F., Gorham, W. A., Orr, D. B., and Goldberg, I. *The talents of American youth: I. Design for a study of American youth.* Boston: Houghton Mifflin, 1962.

Harris, C. W. Statistical methods. In R. L. Ebel (Ed.), *Encyclopedia of educational research* (4th ed.). New York: Macmillan, 1969.

Herriott, R. E. Survey research method. In R. L. Ebel (Ed.), *Encyclopedia of educational research* (4th ed.). New York: Macmillan, 1969.

Isaac, S. and Michael, W. B. *Handbook in research and evaluation*. San Diego, Calif.: EdITS, 1971.

Kerlinger, F. N. Research in education. In R. L. Ebel (Ed.) *Encyclopedia of educational research* (4th ed.). New York: Macmillan, 1969.

Kerlinger, F. N. Foundations of behavioral research (2nd ed.). New York: Holt, Rinehart and Winston, 1973.

Krathwohl, D. R. An analysis of the perceived ineffectiveness of educational research and some recommendations. *Educational Psychologist*, 1974, *11*, 73-86.

Mouly, George J. Research methods. In R. L. Ebel (Ed.) *Encyclopedia of educational research* (4th ed.). New York: Macmillan, 1969.

Smith, B. O. and Meux, M. Research in teacher education: Problems, analysis and criticism. In F. R. Cypherts and E. Spaights (Eds.), *An analysis and projection of research in teacher education*. Washington, D.C.: USOE Cooperative Research Project No. F-015, 1964. (ERIC Document Reproduction Service No. ED 003399.)

Tatsuoka, M. M. Experimental methods. In R. L. Ebel (Ed.), *Encyclopedia of educational research* (4th ed.). New York: Macmillan, 1969.

Travers, R. M. W. Utilization in education of knowledge derived from research on learning. In F. R. Cypherts and E. Spaights (Eds.), *An analysis and projection of research in teacher education*. Washington, D.C.: USOE Cooperative Research Project No. F-015, 1964. (ERIC Document Reproduction Service No. ED 003399.)

Trow, M. Survey research and education. In C. Y. Glock (Ed.), *Survey research in the social sciences*. New York: Russell Sage Foundation, 1967.

West, L. J. Sampling techniques. *Delta Pi Epsilon Journal,* 1963, 5(4), 115-125.

West, L. J. *Acquisition of typewriting skills.* Belmont, Calif.: Pitman/Fearon, 1969.

West, L. J. Survey of entry-level bookkeeping activities in relation to the high school bookkeeping curriculum. *Research Report 73-1.* New York: City University of New York (Institute for Research and Development in Occupational Education), 1973. (ERIC Document Reproduction Service No. ED 086873.) [Extensively summarized in the November 1974 issue of the *Delta Pi Epsilon Journal,* pp. 1-34.]

West, L. J. Implications of research for teaching typewriting (2d ed.). *Delta Pi Epsilon Research Bulletin No. 4,* 1974. The Association (St. Peter, Minnesota: Gustavus Adolphus College).

West, L. J. Principles and procedures for testing of typewriting and stenographic proficiency. *Business Education Forum,* 1975, *29*(5), 24-27.

DELTA PI EPSILON fraternity, founded in 1936, is an honorary graduate organization for men and women devoted to the advancement and professionalization of business education. Through its ideals of service, leadership and cooperation, the fraternity strives to make significant and unique contributions to professional growth and scholarly achievement in business education. In the words of its founder, Dr. Paul Lomax, can be seen the scope of the fraternity, "The professional interests of Delta Pi Epsilon encompass the whole of business education in relation to the entire fields of American business and American education. Its membership... must always think in terms of the common good and advancement of all our business teachers and of all students who pursue courses in business education."

DELTA PI EPSILON CHAPTERS, INSTITUTIONS, CITIES, STATES

ALPHA: New York University, New York, NY, 10003
BETA: Oklahoma State University, Stillwater, OK 74074
GAMMA: University of Pittsburgh, Pittsburgh, PA 15260
DELTA: University of Cincinnati, Cincinnati, OH 45221
EPSILON: Boston University, Boston, MA 02215
ZETA: University of North Carolina, Greensboro, NC 27412
ETA: University of Denver, Denver, CO 80210
THETA: Indiana University, Bloomington, IN 47405
IOTA: Syracuse University, Syracuse, NY 13210
KAPPA: University of Michigan, Ann Arbor, MI 48103
LAMBDA: Northwestern University, Evanston, IL 60201
MU: University of Tennessee, Knoxville, TN 37916
NU: University of Kentucky, Lexington, KY 40506
XI: University of Florida, Gainesville, FL 32603
OMICRON: University of Iowa, Iowa City, IA 52240
PI: Ball State University, Muncie, IN 47306
RHO: Ohio State University, Columbus, OH 43210
SIGMA: University of Oklahoma, Norman, OK 73069
TAU: Columbia University, New York, NY 10027 (inactive)
UPSILON: University of Mississippi, University, MS 38677
PHI: University of Minnesota, Minneapolis, MN 55455
CHI: Pennsylvania State University, State College, PA 16802 (inactive)
PSI: University of Southern California, Los Angeles, CA 90007
OMEGA: George Peabody College for Teachers, Nashville, TN 37203
ALPHA ALPHA: University of Northern Colorado, Greeley, CO 80631
ALPHA BETA: University of Illinois, Urbana, IL 61803
ALPHA GAMMA: University of Houston, Houston, TX 77004
ALPHA DELTA: Emporia State University, Emporia, KS 66801
ALPHA EPSILON: North Texas State University, Denton, TX 76203
ALPHA ZETA: Temple University, Philadelphia, PA 19122
ALPHA ETA: University of Wisconsin—Madison, Madison, WI 53706
ALPHA THETA: University of Texas, Austin, TX 78712
ALPHA IOTA: University of Colorado, Boulder, CO 80304
ALPHA KAPPA: San Francisco State University, San Francisco, CA 94132
ALPHA LAMBDA: Michigan State University, East Lansing, MI 48823
ALPHA MU: State University of New York, Albany, NY 12203
ALPHA NU: University of North Dakota, Grand Forks, ND 58202
ALPHA XI: Hunter College of the City University of New York, New York, NY 10021
ALPHA OMICRON: University of California at Los Angeles, Los Angeles, CA 90024
ALPHA PI: Wayne State University, Detroit, MI 48202
ALPHA RHO: California State University, Fresno, Fresno, CA 93726
ALPHA SIGMA: Arizona State University, Tempe, AZ 85281
ALPHA TAU: University of Northern Iowa, Cedar Falls, IA 50631
ALPHA UPSILON: University of Nebraska, Lincoln, NE 68508
ALPHA PHI: Northern Illinois University, DeKalb, IL 60115
ALPHA CHI: Rider College, Lawrenceville, NJ 08648
ALPHA PSI: Mankato State University, Mankato, MN 56001
ALPHA OMEGA: Brigham Young University, Provo, UT 84601
BETA ALPHA: Indiana University of Pennsylvania, Indiana, PA 15701
BETA BETA: Southern Illinois University at Edwardsville, Edwardsville, IL 62025
BETA GAMMA: Virginia Polytechnic Institute and State University, Blacksburg, VA 24061
BETA DELTA: University of Georgia, Athens, GA 30601
BETA EPSILON: San Jose State University, San Jose, CA 95192
BETA ZETA: Indiana State University, Terre Haute, IN 47809
BETA ETA: Bowling Green State University, Bowling Green, OH 43403
BETA THETA: University of Wisconsin—Whitewater, Whitewater, WI 53190
BETA IOTA: Illinois State University, Normal, IL 61761
BETA KAPPA: Portland State University, Portland, OR 97207
BETA LAMBDA: Shippensburg State College, Shippensburg, PA 17257
BETA MU: Central Connecticut State College, New Britain, CT 06050
BETA NU: Utah State University, Logan, UT 84321
BETA XI: Memphis State University, Memphis, TN 38111
BETA OMICRON: Southern Illinois University at Carbondale, Carbondale, IL 62901
BETA PI: California State University, Los Angeles, Los Angeles, CA 90032
BETA RHO: Western Michigan University, Kalamazoo, MI 49001
BETA SIGMA: University of Wisconsin—Eau Claire, Eau Claire, WI 54701
BETA TAU: Georgia State University, Atlanta, GA 30303
BETA UPSILON: Pittsburg State University, Pittsburg, KS 66762
BETA PHI: Montclair State College, Upper Montclair, NJ 07043
BETA CHI: Western Illinois University, Macomb IL 61455
BETA PSI: Eastern Illinois University, Charleston, IL 61920
BETA OMEGA: Louisiana Tech University, Ruston, LA 71270
GAMMA ALPHA: Eastern Michigan University, Ypsilanti, MI 48197
GAMMA BETA: Trenton State College, Trenton, NJ 08625
GAMMA GAMMA: Virginia Commonwealth University, Richmond, VA 23284
GAMMA DELTA: University of Rhode Island, Kingston, RI 02881